Policy Gradient

虎尾科技大學
40723115
林于哲

January 16 2021

Contents

1 Definition

π : policy
s : States
a : Actions
r : Rewards
S_t, A_t, R_t : State,Action and Reward at time step 't' of one trajectory
γ : Discount Factor; 懲罰不確定的未來 reward

G_t : Return;Discounted future reward$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

$P(s', r|s, a)$: 伴隨著現在的 a 和 r 的 state 前往下一個 state’s’ 的轉移機率矩陣 (單階)

$\pi(a|s)$: 隨機策略 (agent 的行為策略)

$\pi_\theta(.)$: 被 θ 參數化的策略

$\mu(s)$: 確定的策略;we also lable this as $\pi(s)$ using a different letter gives better distinction so thst we can easily tell when the policy is stochastic or deterministic

$V(s)$: ’ 狀態值函數’ 測量 state 的預期收益 (報酬率)

$V^\pi(s)$: 根據 policy 的狀態值函數 $V^\pi(s) = \mathbb{E}_{a\sim\pi}[G_t|S_t = s]$

$Q(s, a)$: ’ 行為值函數’ 評估一對 state and action 的預期收益

$Q_w(.)$: 被 w 參數化的行為值函數

$Q^\pi(s, a)$: 根據 policy 的行為值函數 $Q^\pi(s, a) = \mathbb{E}_{a\sim\pi}[G_t|S_t = s, A_t = a]$

$A(s, a)$: Advantage Function,$A(s, a) = Q(s, a) - V(s)$; 像是另一種版本的 Q-value; 由狀態值為基準降低方差

參數化: 待軟體建置於一給定環境時，再依該環境的實際需求填選參數，即可成為適合該環境的軟體。

The goal of reinforcement learning : Find an optimal behavior strategy for the agent to obtain optimal reward

The goal of policy gradient : Modeling and optimizing the policy directly

The value of reward function : 取決於策略，克應用各種算法 optimizeθ，已獲得最佳 reward

defined as:

$$J(\theta) = \sum_{s\in\mathcal{S}} d^\pi(s)V^\pi(s) = \sum_{s\in\mathcal{S}} d^\pi(s) \sum_{a\in\mathcal{A}} \pi_\theta(a|s)Q^\pi(s, a)$$

2 Mokov chain

stochastic process : 將隨著時間變化的狀態，以數學模式表示

$+$

Mokov property : 在目前以及所有過去事件的條件下，任何未來事件發生的機率，和過去的事件不相關僅和目前狀態相關 $d^\pi(s)$: π_θ 的 Mokov chain 的平穩分布 (在 π 下的策略狀態分佈) $d^\pi(s) = \lim_{t\to\infty} P(s_t = s|s_0, \pi_\theta)$

$*$ 當策略在其他函數的下標時，將省略 π_θ 的 θ e.g. $d^\pi(s)$ and Q^π should be $d^{\pi\theta}(s)$、$Q^{\pi\theta}$

stationary probability for π_θ : 隨著時間進展，結束一個狀態保持不變的機率分布

3 為什麼不用 value-base 而是 policy-base

因為要估計其值得動作和狀態數不勝數，因此在連續空間計算成本太高，θ 向 $\nabla_\theta J(\theta)$ 建議方向移動，已找到 $\pi\theta$ 的最佳 θ，從而產生最高回報。

4 Proof Policy Gradient Theorem

計算 $\nabla_\theta J(\theta)$depends on 動作選擇和目標選擇行為之後狀態的靜態分布，而導致計算困難。
Policy gradient theorem : 為目標函式的導數重新建構，使它不涉及 $d^\pi(.)$ 的導數。

$$\nabla_\theta J(\theta) = \nabla_\theta \sum_{s\in\mathcal{S}} d^\pi(s) \sum_{a\in\mathcal{A}} Q^\pi(s,a)\pi_\theta(a|s) \propto \sum_{s\in\mathcal{S}} d^\pi(s) \sum_{a\in\mathcal{A}} Q^\pi(s,a)\nabla_\theta\pi_\theta(a|s)$$

$$\nabla_\theta J(\theta) = \nabla_\theta V^\pi(s_0) = \sum_{s\in\mathcal{S}} d^\pi(s) \sum_{a\in\mathcal{A}} \pi_\theta(a|s)Q^\pi(s,a)\frac{\nabla_\theta\pi_\theta(a|s)}{\pi_\theta(a|s)}$$
$$= \mathbb{E}_\pi[Q^\pi(s,a)\nabla_\theta \ln\pi_\theta(a|s)]_{;Because(\ln x)'=1/x}$$

$*$ 當狀態和動作分布都遵循策略 π_θ 時 $\mathbb{E}_\pi$ 表示 $\mathbb{E}_{\pi\sim d\pi, a\sim\pi_\theta}$

Proof $\nabla_\theta J(\theta = \nabla_\theta V^\pi(s_0) \quad \nabla_\theta V^\pi(s)$

$=\nabla_\theta\Big(\sum_{a\in\mathcal{A}} \pi_\theta(a|s)Q^\pi(s,a)\Big)$

$= \sum_{a\in\mathcal{A}}\Big(\nabla_\theta\pi_\theta(a|s)Q^\pi(s,a) + \pi_\theta(a|s) red\nabla_\theta Q^\pi(s,a)\Big)_{;Derivative product rule.}$

$= \sum_{a\in\mathcal{A}}\Big(\nabla_\theta\pi_\theta(a|s)Q^\pi(s,a)+\pi_\theta(a|s)red\nabla_\theta \sum_{s',r} P(s',r|s,a)(r+V^\pi(s'))\Big)_{;Extend Q^\pi with future state value.}$

$= \sum_{a\in\mathcal{A}}\Big(\nabla_\theta\pi_\theta(a|s)Q^\pi(s,a) + \pi_\theta(a|s)red\sum_{s',r} P(s',r|s,a)\nabla_\theta V^\pi(s')\Big)_{P(s',r|s,a) or r is not a func of \theta}$

$= \sum_{a\in\mathcal{A}}\Big(\nabla_\theta\pi_\theta(a|s)Q^\pi(s,a) + \pi_\theta(a|s)red\sum_{s'} P(s'|s,a)\nabla_\theta V^\pi(s')\Big)_{;Because P(s'|s,a)=\sum_r P(s',r|s,a)}$

Let $\phi(s) = \sum_{a\in\mathcal{A}} \nabla_\theta\pi_\theta(a|s)Q^\pi(s,a)$; 當 K = 1，我們把所有可能動作總結到目標狀態的轉移機率 $\rho^\pi(s\to x,k+1) = \sum_{s'}\rho^\pi(s\to s',k)\rho^\pi(s'\to x,1)$

$red\nabla_\theta V^\pi(s)$

$= \phi(s) + \sum_a \pi_\theta(a|s)\sum_{s'} P(s'|s,a) red\nabla_\theta V^\pi(s')$

$= \phi(s) + \sum_{s'}\sum_a \pi_\theta(a|s)P(s'|s,a)red\nabla_\theta V^\pi(s')$

$= \phi(s) + \sum_{s'}\rho^\pi(s\to s',1)red\nabla_\theta V^\pi(s')$

$= \phi(s) + \sum_{s'}\rho^\pi(s\to s',1)red\nabla_\theta V^\pi(s')$

$= \phi(s) + \sum_{s'}\rho^\pi(s\to s',1)red[\phi(s') + \sum_{s''}\rho^\pi(s'\to s'',1)\nabla_\theta V^\pi(s'')]$

$= \phi(s) + \sum_{s'}\rho^\pi(s\to s',1)\phi(s') + \sum_{s''}\rho^\pi(s\to s'',2)red\nabla_\theta V^\pi(s'')_{;Consider s' as the middle point for s\to s''}$

$= \phi(s) + \sum_{s'}\rho^\pi(s\to s',1)\phi(s') + \sum_{s''}\rho^\pi(s\to s'',2)\phi(s'') + \sum_{s'''}\rho^\pi(s\to s''',3)red\nabla_\theta V^\pi(s''')$

$= \ldots_{;Repeatedly unrolling the part of \nabla_\theta V^\pi(.)}$

$= \sum_{x\in\mathcal{S}}\sum_{k=0}^{\infty}\rho^\pi(s\to x,k)\phi(x)$

$\nabla_\theta J(\theta) = \nabla_\theta V^\pi(s_0)_{;Startingfromarandomstates_0}$
$= \sum_s blue \sum_{k=0}^{\infty} \rho^\pi(s_0 \to s, k)\phi(s)_{;Letblue\eta(s)=\sum_{k=0}^{\infty} \rho^\pi(s_0 \to s,k)}$
$= \sum_s \eta(s)\phi(s) = \left(\sum_s \eta(s)\right) \sum_s \frac{\eta(s)}{\sum_s \eta(s)} \phi(s)_{;Normalize\eta(s),s\in\mathcal{S}tobeaprobabilitydistribution.}$
$\propto \sum_s \frac{\eta(s)}{\sum_s \eta(s)} \phi(s)_{\sum_s \eta(s)isaconstant}$
$= \sum_s d^\pi(s) \sum_a \nabla_\theta \pi_\theta(a|s) Q^\pi(s,a)_{d^\pi(s)=\frac{\eta(s)}{\sum_s \eta(s)}isstationarydistribution.}$

$\nabla_\theta J(\theta) \propto \sum_{s\in\mathcal{S}} d^\pi(s) \sum_{a\in\mathcal{A}} Q^\pi(s,a)\nabla_\theta \pi_\theta(a|s)$
$= \sum_{s\in\mathcal{S}} d^\pi(s) \sum_{a\in\mathcal{A}} \pi_\theta(a|s) Q^\pi(s,a) \frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)}$
$= \mathbb{E}_\pi[Q^\pi(s,a)\nabla_\theta \ln \pi_\theta(a|s)]_{;Because(\ln x)'=1/x}$

5 Policy Gradient Theorem

Policy Gradient 通過反覆估計梯度來最大化預期的總 reward
$g = \nabla_\theta \mathbb{E}[\sum_{t=0}^{\infty} r_t]$; $g = \mathbb{E}[\sum_{t=0}^{\infty} \psi_t \nabla_\theta log \pi_\theta(a_t|s_t)]$
ψ_t may be one of following:

- $\sum_{t=0}^{\infty} r$: total reward of trajectory
- $\sum_{t'=t}^{\infty} r'$: reward following action a_t
- $\sum_{t'=t}^{\infty} r'_t - b(s_t)$: baseline version of previous formula
- $Q^\pi(s_t, a_t)$: state-action value function
- $A^\pi(s_t, a_t)$: Advantage Function
- $r_t + V^\pi(s_t + 1) - V^\pi(s_t)$: TD residual

The letter formulas use the definitions
$V^\pi(s_t) = \mathbb{E}_{st+1:\infty, at:\infty}[\sum_{l=0}^{\infty} r_t + l]$
$Q^\pi(s_t, a_t) = \mathbb{E}_{st+1:\infty, at+1:\infty}[\sum_{l=0}^{\infty} r_t + l]$
$A^\pi(s_t, a_t) = Q^\pi(s_t, a_t) - V^\pi(s_t)(AdvantageFunction)$

6 Actor Crtic

原始的 policy gradient 沒有 bias，但方差大;Many following algorithms were proposed to reduce variance while keeping thebias unchanged

$$g = \mathbb{E}[\sum_{t=0}^{\infty} \psi_t \nabla_\theta log \pi_\theta (a_t|s_t)]$$

Actor-Critic : reduce gradient variance in vanilla policy consist of two models

Critic : updates the value function parameter w and depending on the algorithm it could be action-value $Q_w(a|s)$or state-value $V_w(s)$

Actor : update the policy parameters θ for $\pi_\theta(a|s)$,in direction suggested by critic

How it work in a simple action-value actor-critic:

- Initialize s,θ,w at random;sample a $\sim \pi_\theta(a|s)$
- For $t = 1 \sim T$:
  1 Sample reward r$_t$$\sim R(s,a)$ and next state $s^{'} \sim P(s^{'}|s,a)$
  2 The sample the next action $a^{'} \sim \pi_\theta(a^{'}|s^{'})$
  3 Update the policy parameters θ :
  $$\theta \leftarrow \theta + \alpha_\theta Q_w(s,a) \nabla_\theta ln \pi_\theta(a|s)$$
  4 Compute the correction (TD error) for action-value at time t:
  $$\delta = r_t + \gamma Q_w(s^{'},a^{'}) - Q_w(s,a)$$
  and use it to update the parameters of action - value function:
  $$w \leftarrow w + \alpha_w \delta \nabla_w Q_w(s,a)$$
  5 Update $a \leftarrow a^{'} ands \leftarrow s^{'}$; learning rate : a_θ and a_w